

# **GNN: Графовые нейронные сети**

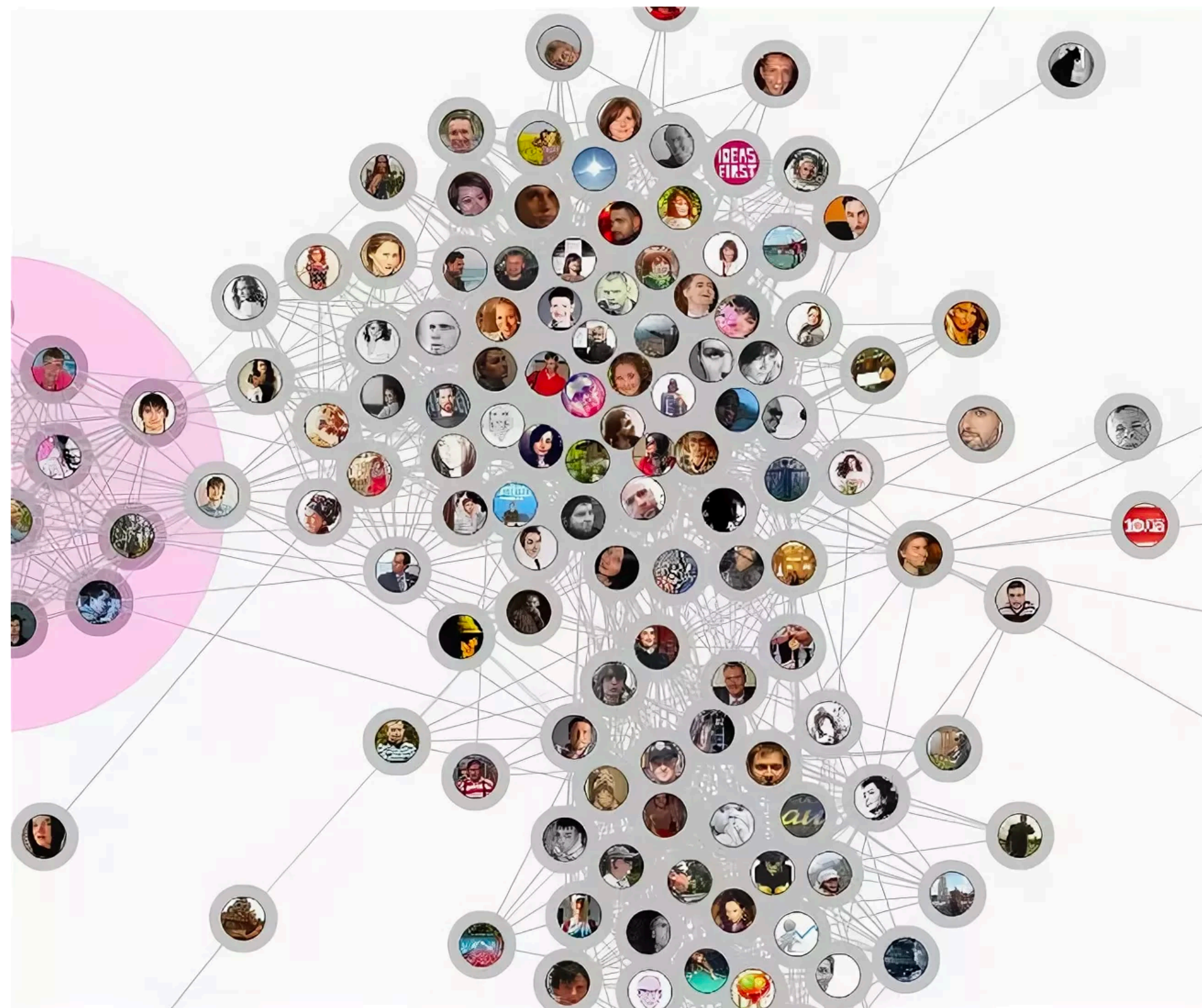
# План:

1. Введение
2. Недостатки существующих подходов
3. Основные задачи на графах
4. Постановка задачи
5. Пример использования GNN для предсказания тренда акций

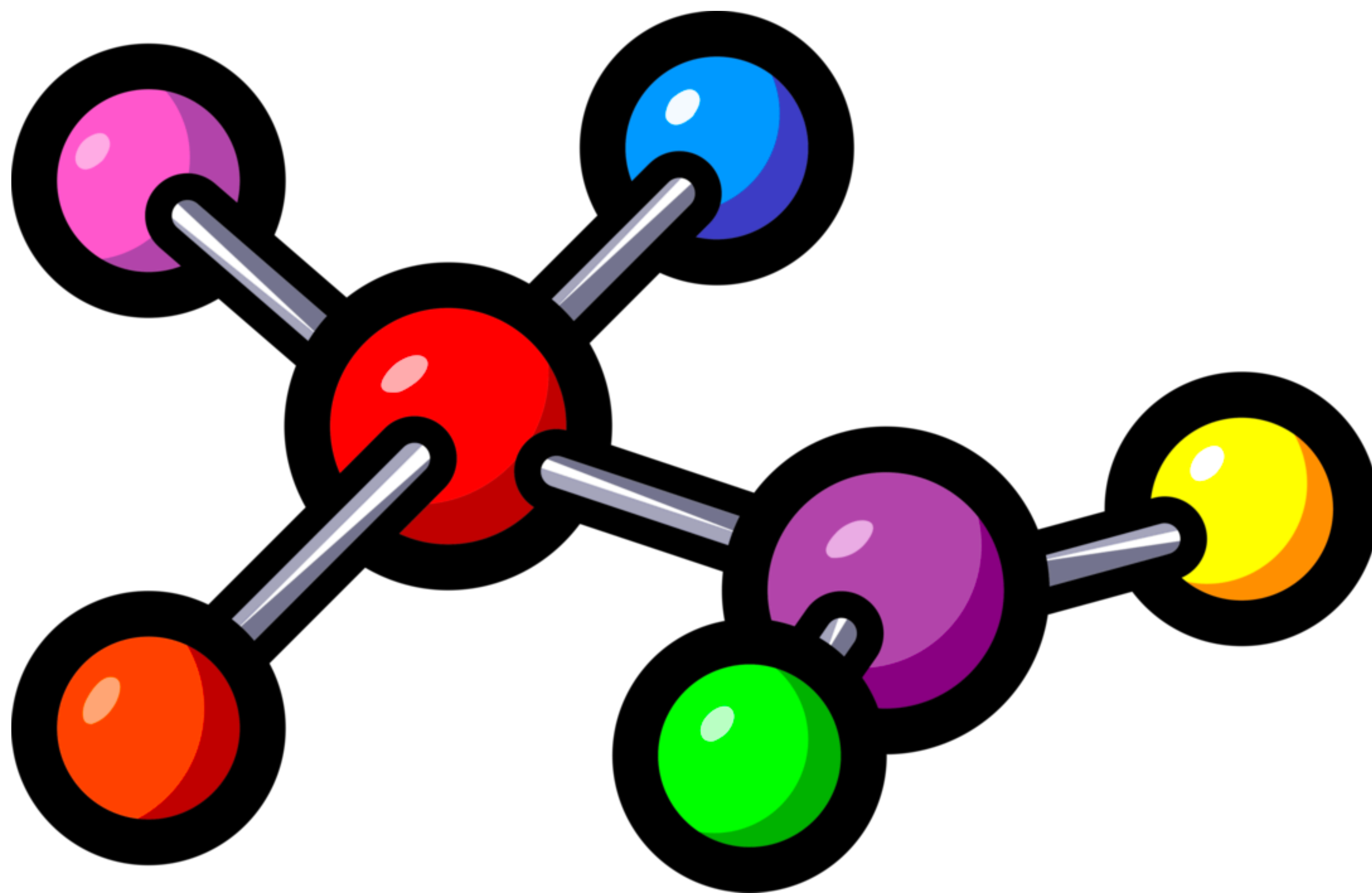
# Введение

Многие задачи современной науки о данных решают задачи работы с данными, структурированными особым образом. Часто между отдельными элементами данных могут быть установлены некоторые взаимные отношения. Примерами таких данных могут быть:

# Введение

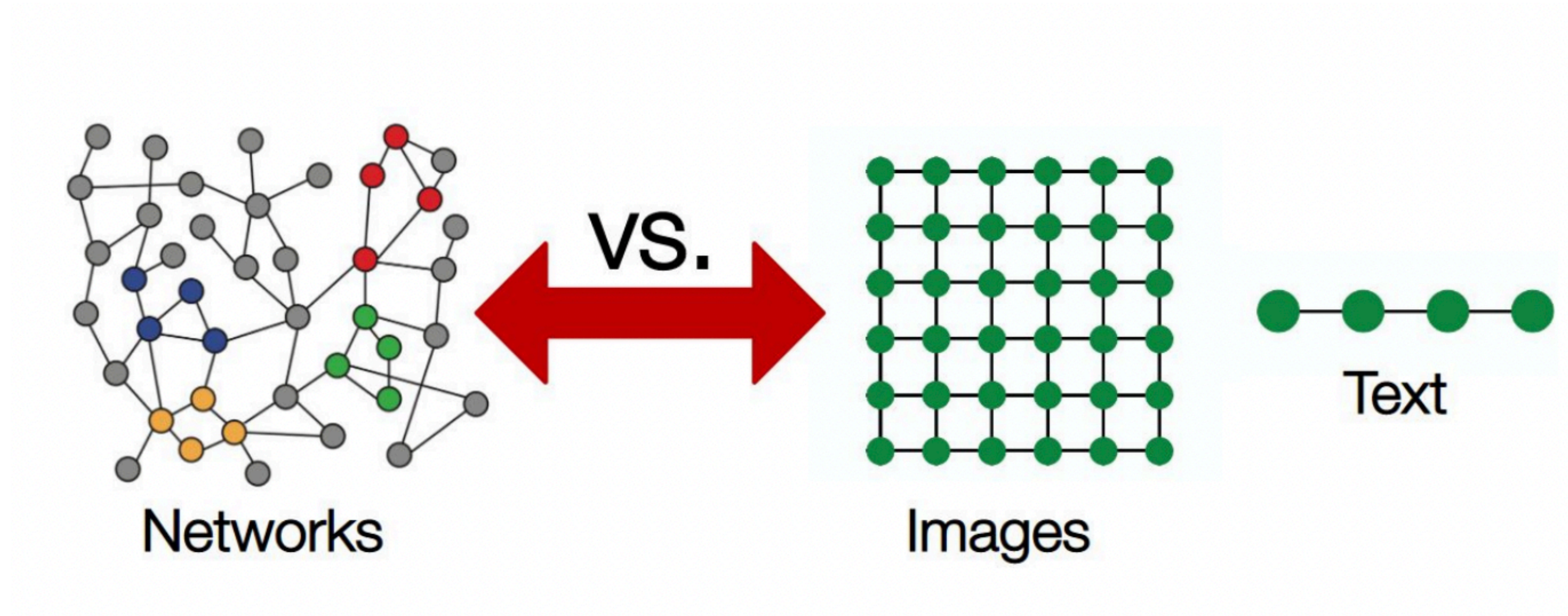


# Введение





# Недостатки существующих подходов



# Основные задачи на графах

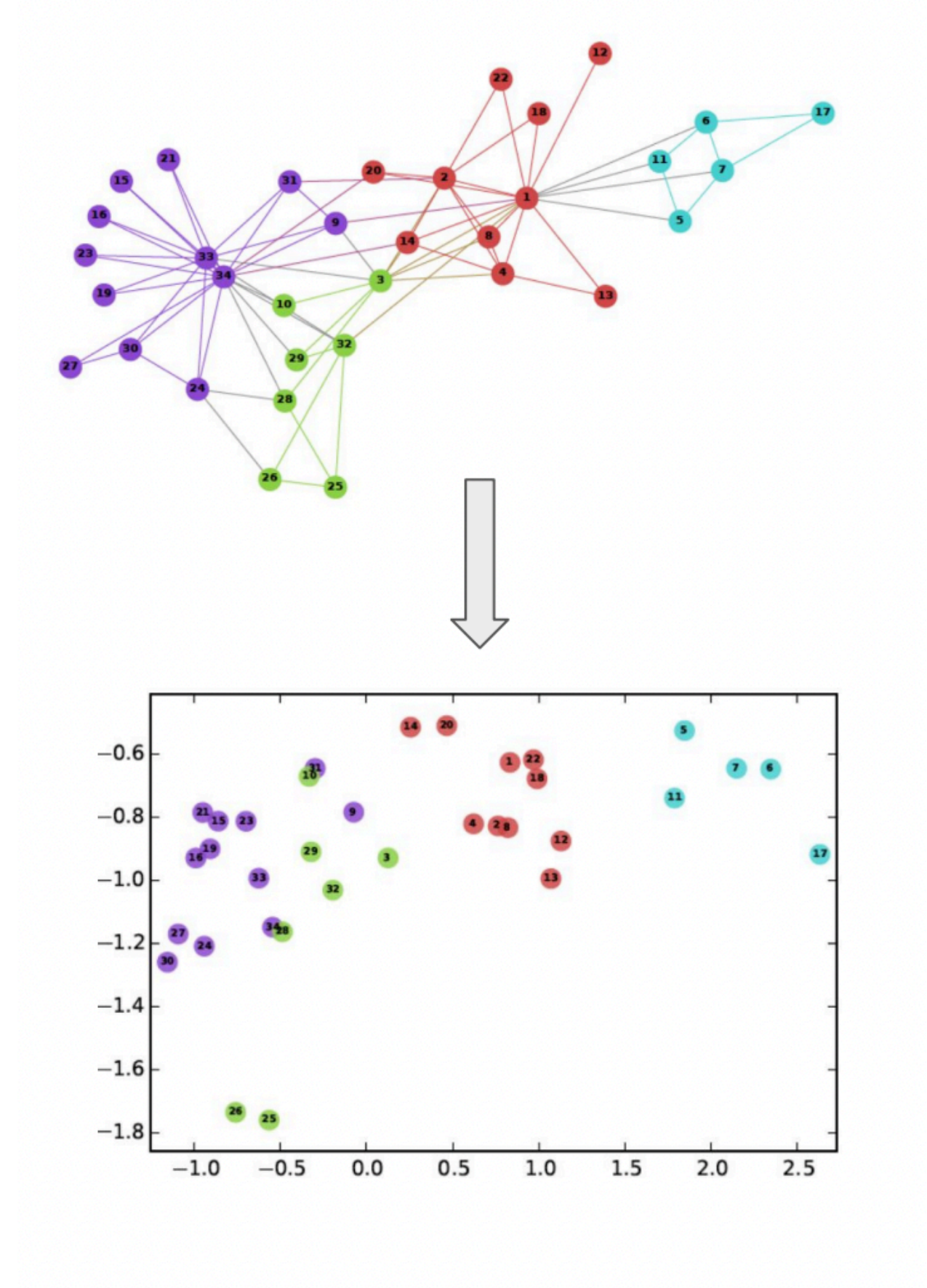
С учителем, частично с учителем	С учителем, частично с учителем	Без учителя
Классификация	Предсказание связи	Детекция сетей
<ul style="list-style-type: none"><li>- является ли аккаунт ботом?</li><li>- предсказание возраста/пола/</li><li>- предсказание функции нового белка на основе его взаимодействий с другими</li><li>- предсказание тематики статьи на основе ее цитирований</li></ul>	<ul style="list-style-type: none"><li>- предсказание взаимодействий лекарств</li><li>- Построение рекомендательной системы</li></ul>	<ul style="list-style-type: none"><li>- выявление групп белков</li><li>- Выявление сформировавшихся подсетей социального графа</li></ul>



# Формальная постановка задачи

Задача:

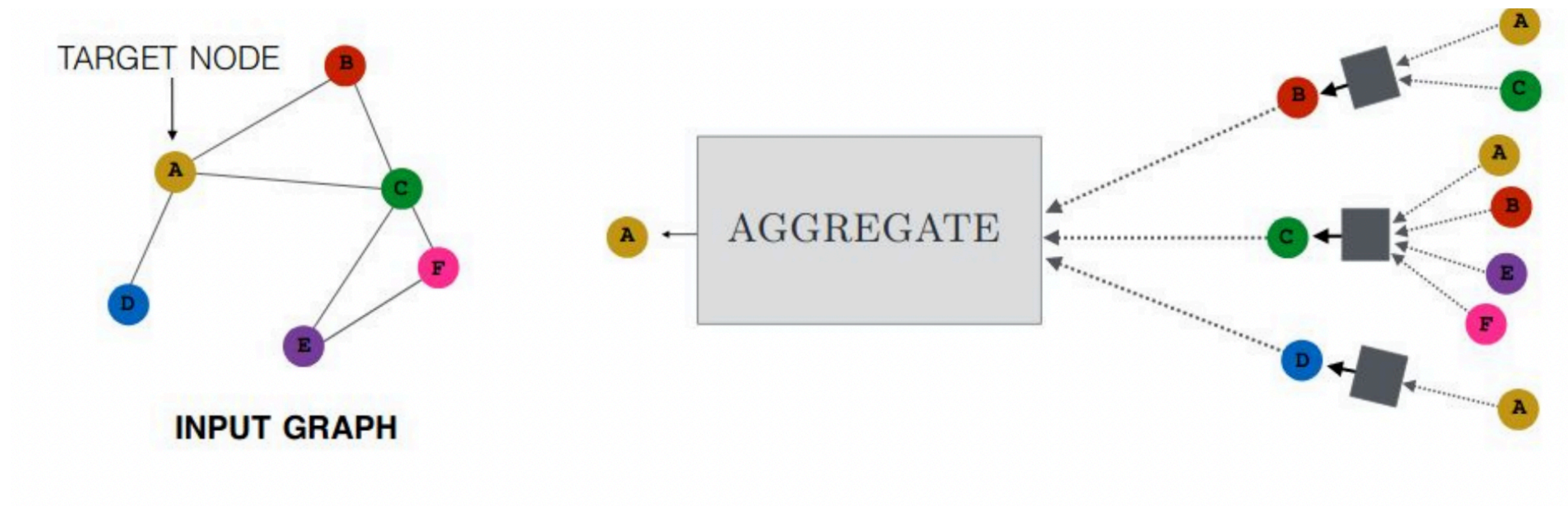
Найти такое представление вершин заданного графа в виде некоразмерного векторного пространства, которое сохраняет наибольшее количество информации как о самих вершинах в отдельности, так и об их взаимных отношениях.





# Основная идея GNN

Основная идея графовых нейронных сетей состоит в формировании так называемых UPDATE и AGGREGATE шагов для построения низкоразмерного признакового пространства.



# Задача предсказания тренда акций



# Задача предсказания тренда акций

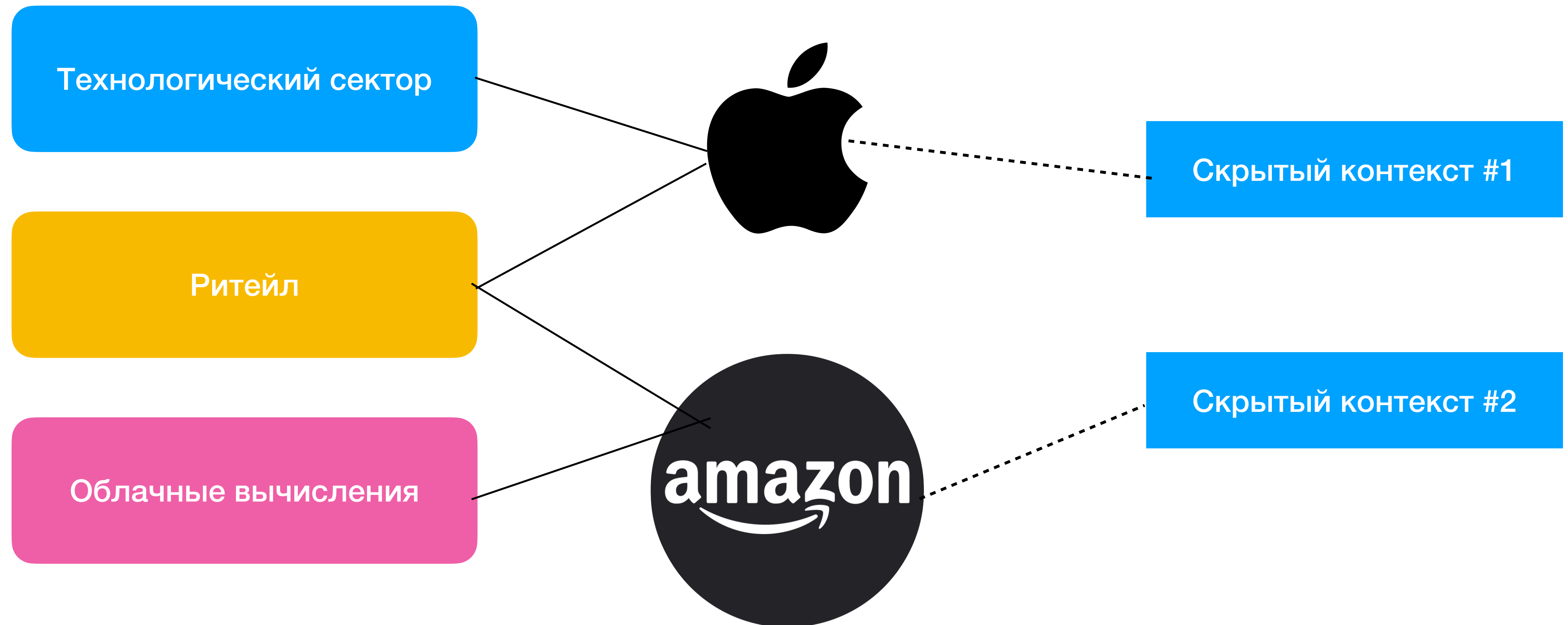
Формальное  
представление целевой  
переменной

$$d_i^t = \frac{\text{Price}_i^{t+1} - \text{Price}_i^t}{\text{Price}_i^t}$$

Функционал ошибки

$$\mathcal{L} = \sum_{t \in \mathcal{T}} \text{MSE} (p^t, d^t) = \sum_{t \in \mathcal{T}} \sum_{i \in \mathcal{S}^t} \frac{(p_i^t - d_i^t)^2}{|\mathcal{S}^t|}$$

# Задача предсказания тренда акций





# Метрики качества

1. Information Coefficient

**IC**

2. Rank Information Coefficient

3. Precision at {3,5,10,30}

$$r_{xy} = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2 \sum_{i=1}^m (y_i - \bar{y})^2}} = \frac{\text{cov}(x, y)}{\sqrt{s_x^2 s_y^2}}$$

# Иные подходы для сравнения

1. Полносвязная MLP модель, 300 миллионов параметров
2. LSTM based модель, 50 миллионов параметров

# Результаты

	MLP	LSTM	GNN
IC	0.0071	0.0097	0.120
Rank IC	0.0061	0.0091	0.115
Precision @ 10	0.5549	0.5904	0.6107

# Результаты

CR стратегии

